Final Project Report - Bank Customer Churn (Group 5)

Soo Young Lee, Ria Kalluri, Cia Cui, Pia Gajjar, Caomengyu Xue, Anya Rajan

Introduction: Customer churn is a very critical issue for banks to address, particularly when retaining existing customers is much more cost-effective than acquiring new ones. In this project, we analyze bank customer churn using machine learning techniques to develop predictive models which will help banks identify at-risk customers early as well optimize retention strategies. Our report will cover exploratory data analysis (EDA), feature engineering, and multiple machine learning models, including Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, MLP Classifier, XGBoost, LightGBM, and CatBoost to find the most effective model for churn prediction.

Bank customer churn occurs when customers close their accounts, leading to revenue loss and increased customer acquisition costs for the business. Identifying churn early allows banks to take proactive measures to improve both their customer satisfaction and retention. Our project's aim is to build a reliable churn prediction model that balances accuracy, recall, and precision, ensuring effective intervention strategies. Through data-driven analysis and model comparisons, we will identify the most effective model for predicting churn and also extract key factors that drive customer attrition. These insights will ultimately help the bank implement better targeted retention strategies, reduce unnecessary retention and acquisition costs, and improve overall customer loyalty.

EDA/Feature Engineering: We first checked for missing values and duplicates, confirming a complete dataset. *CreditScore* follows a near-normal distribution, while *Age* is slightly right-skewed, with most customers aged 30-40. *Balance* has many zero values, likely indicating inactive accounts, while other balances cluster around 100K-150K. *EstimatedSalary* is uniformly distributed, suggesting minimal influence on churn prediction.

For categorical features, *Geography* shows higher churn rates in Germany, while France and Spain retain more customers. *Gender* analysis indicates women churn more, possibly due to banking service sensitivity. *Activity status* is a key factor, with inactive customers showing significantly higher churn. *Card Type* has minimal impact on churn. Boxplots confirmed no significant outliers in *Balance* and *EstimatedSalary*. Given the class imbalance (20% churn, 80% retention), we initially used unbalanced data, which improved accuracy but failed to identify churned customers. Since missing churned customers is costlier than false positives, we prioritized recall and used balanced data.

To ensure fair contribution of numerical features, we applied standard scaling, preventing the model from overemphasizing *Balance*. For categorical variables, One-Hot Encoding transformed non-numeric features like *Geography* into binary variables (e.g., *Geography_Spain*, *Geography_Germany*), enhancing interpretability. The dataset was split into 80% training and 20% testing. We initially experimented with

Logistic Regression, Random Forest, Gradient Boosting, and MLP, later focusing on XGBoost, LightGBM, and CatBoost for improved predictive performance.

Model 1 Logistic Regression: To accurately predict bank customer churn, the first model that we evaluated was the logistic regression model. We believe that balancing accuracy, recall, and precision maximizes business impact, hence those were the factors we strongly focused on. Each customer is given a probability of churning, known as the threshold, and it is by default set at 0.5 but we computed in the second model that the optimal threshold for this model would be 0.47. Initially, our baseline model (Threshold = 0.50) achieved 72.82% accuracy, with 72% recall and 73% precision, and effectively distinguished churners from non-churners with an ROC AUC of 0.7862. However, it missed 111 actual churners, posing a risk of revenue loss due to undetected at-risk customers. To address this, we changed the threshold to 0.47, increasing recall to 75% and precision to 74%. Accuracy dropped slightly to 72.69%, and overall the model captured more churners at the cost of a slight rise in false positives (116 vs. 101). Additionally, one thing we would like to point out is that balancing the dataset reduced the total sample size, which explains why the confusion matrix values are lower. Despite this, recall and precision remained strong, ensuring the model effectively identified churners without bias toward non-churners.

Model 2 Random Forest Classifier: For our second model predicting bank customer churn, we began with 18 features (a mix of categorical and numerical) and reduced them to 15 by dropping columns with excessive unique values (RowNumber, CustomerId, Surname). Using GridSearchCV, we tuned max_features [4,7,10,13,15] and max_depth [4,7,10,13,None] to balance feature selection and tree complexity, with the optimal configuration being max_depth = 4 and max_features = 4. The model achieved 72.7% accuracy (the model correctly predicted most customers) with strong precision (74.8%, the model is confident in its churn predictions) and ROC AUC (80.2%, the model effectively distinguishes between churners and non-churners), indicating reliable churn prediction. However, its recall (69.2%) suggests it misses some churners, which could impact customer retention efforts. The F1-score (71.7%) reflects a balance between precision and recall, but improving recall could help the bank take proactive action to reduce churn. If the primary goal is to maximize retention, prioritizing recall in future models may be beneficial, while still maintaining strong overall performance across key metrics. Overall, this model performs well among all metrics, but not high enough to rely heavily on its results. Later models will highlight better examples of models with strong and correct predictive models.

Model 3 Gradient Boosting Classifier: We analyzed the data using the GradientBoostingClassifier for our third model, which includes two versions: gbc1 and gbc2. For gbc1, we applied GridSearchCV to tune three parameters: n_estimators, max_depth, and learning rate. For gbc2, we optimized the n_iter_no_change

parameter. With the default threshold of 0.5, both models displayed decent accuracy and precision, indicating that they were effective at identifying customers at risk of churning. This suggests that the bank could focus its retention efforts on these customers to maximize the impact of their investment in retention strategies. However, both models exhibited relatively low recall scores compared to accuracy and precision. This indicates that they still missed a significant portion of churners, leading to the risk of losing customers without intervention.

To address this, we adjusted the decision threshold to optimize the F1 score, aiming for a better balance between recall and precision. For gbc1, the threshold was set to 0.424, and for gbc2, it was adjusted to 0.394. This resulted in a slight decrease in accuracy and a trade-off between precision and recall. The models became more sensitive in identifying churners, but this also meant that the bank might allocate retention resources to customers who were less likely to churn. Despite this, the increased recall score meant that the models were more capable of capturing potential churners, reducing the risk of losing customers, which is particularly costly in the banking industry. The F1 scores improved, reflecting a better balance between precision and recall. The ROC-AUC scores remained stable at 0.804 for gbc1 and 0.794 for gbc2, suggesting that the models still demonstrated a strong ability to differentiate between churners and non-churners.

Model 4 MLP Model: The Multi-Layer Perceptron (MLP) model was selected as the optimal model for predicting customer churn due to its balanced performance across key metrics. It achieved an accuracy of 76% and a recall of 73% for churners, indicating its strong ability to identify customers likely to leave. The ROC-AUC score of 0.83 and PR-AUC score of 0.84 further validate its effectiveness in distinguishing between churners and non-churners. In developing this model, we performed extensive feature engineering, including creating interaction variables like Age_Balance and Tenure_Products to capture deeper relationships in customer behavior. The data was balanced using oversampling techniques, and hyperparameters such as learning rate, batch size, and dropout rates were optimized to prevent overfitting while maintaining high recall. Despite these improvements, the model still exhibited a slight bias towards non-churners, indicating room for further refinement.

Model 5 XGBoost/LightGBM/CatBoost: XGBoost, LightGBM, and CatBoost were evaluated for customer churn prediction, with all three models demonstrating similar performance in accuracy (~72%) and AUC (~0.80), indicating their ability to distinguish churned from non-churned customers. However, recall was relatively low (~0.69–0.70) before threshold tuning, meaning a significant portion of actual churners were not being identified. Among the three, CatBoost showed the highest recall (0.70) and F1 Score (0.717), making it slightly more effective at balancing precision and recall.

After threshold tuning, recall improved substantially, with CatBoost increasing to 0.82, ensuring better identification of churned customers. F1 Score also improved across all models, with CatBoost reaching 0.75, showing a better trade-off between recall and precision. As expected, precision decreased slightly due to an

increase in false positives, but this trade-off is justified in churn prediction, where capturing more at-risk customers is the priority. Notably, AUC remained stable (~0.80) before and after threshold tuning, confirming that overall model discrimination was not compromised. After adjustment, CatBoost emerged as the best performer, offering the highest recall and F1 Score, while XGBoost and LightGBM also maintained strong predictive capabilities. The improved recall ensures that more churned customers are correctly identified, making these models more effective in customer retention strategies.

Best Model/Conclusion: We selected MLP (Multi-Layer Perceptron) as our optimal model, as it was the most balanced in terms of performance across key metrics, making sure no single metric dominates at the expense of others. It achieved 76% accuracy (Indicates a strong overall classification performance), 73% recall for churners (Shows a reasonable ability to detect customers likely to leave), 76% recall (Ensures that a good portion of churners and non-churners are correctly identified), 76% precision (Confirms that most predicted churners are actually churners), and 76% F1-score (Reflects a well-balanced trade-off between precision and recall). Additionally, the 0.83 ROC-AUC score demonstrates strong ability to distinguish between churners and non-churners, making it a consistent and well-rounded choice for predicting customer churn.

Final Recommendations: To further enhance the MLP model and align it with business goals, we propose three key strategies. First, optimizing decision thresholds based on business needs—lower thresholds maximize churn detection (recall), while higher thresholds improve precision to reduce intervention costs. Second, enhancing model performance through hyperparameter tuning, leveraging ensemble methods (such as combining MLP with boosting models), and refining feature engineering to capture crucial churn indicators. Lastly, continuous monitoring and adaptation is essential, ensuring the model is regularly retrained with updated data and adjusted based on market trends and customer behaviors. These recommendations will help businesses retain more customers, optimize resource allocation, and make data-driven decisions that balance churn prevention and operational efficiency.